International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN (P): 2249–6831; ISSN (E): 2249–7943
Vol. 12, Issue 1, Jun 2022, 19–28

PREDICTION OF DIABETES USING RANDOM FOREST ANDXGBOOST CLASSIFIERS

ANKUSH GHOSH

Department of Computer Science and Engineering, Acharya Institute of Technology

ABSTRACT

Diabetes is a long-term sickness, which is a category of metabolic disorders caused by a persistently elevated blood sugar level. If accurate early diagnosis is feasible, diabetes's and severity and risk factors may be greatly decreased. The small amount of labelled data and the emergence of outliers (missing data) in diabetes datasets makes prediction of diabetes reliably and accurately more difficult. In this paper, we accepted a rigorous model for predicting diabetes that includes outlier elimination, missing value filling, data standardization, function collection, cross-validation, and a variety of Machine Learning classifiers (Random Forest and XGBoost).

The performance metric, which is then maximised using the grid search technique during hyperparameter tuning. Using the PID Dataset, all of the studies in this literature were carried out under the same test conditions. Our proposed classifier outperforms the Random Forest classifier by 3.14% with an Accuracy of 0.750. Our suggested model outperforms the other approaches discussed in the paper for diabetes prediction. It produces better results in the same dataset, resulting in better diabetes prediction accuracy. Our diabetes prediction source code has been made public.

KEYWORDS: Diabetes Prediction, Pima Indian Diabetic Dataset, Machine Learning, Missing Values & Outliers

Received: Nov 28, 2021; **Accepted**: Dec 18, 2021; **Published**: Jan 04, 2022; **Paper Id**: IJCSEITRJUN20223

Original Article

1. INTRODUCTION

In today's world, diabetes is a very well-known phrase that causes substantial issues in both developed and poor countries. Insulin is the hormone that the pancreas produces, that causes glucose to enter the circulatory system from a meal. Diabetes happens because of lack of the hormone which is created by pancreatic breakdown, and triggersretinal impairment, paralysis and renal, cerebral vascular dysfunction, cardiovascular dysfunction, joint weakness, pathogenic immune response, peripheral vascular diseases and weight loss[1].

While there is no long-term solution to diabetes, it is managed and avoided if a correct early diagnosis is possible. Its estimation is difficult since the distribution of groups for all characteristics is not separable along the line.

Between 1980 and 2014, the prevalence of diabetes in adults (aged 18 or above) grew from 4.7 % to 8.5 %, according to research on diabetes patients, and is increasingly increasing in second and third world countries[2]. According to statistics from 2017, Diabetes affects 451 million people globally, with that figure predicted to grow to 693 million till 2045.[3].Another scientific analysis revealed the prevalence of diabetes, reporting that half a billion people globally had diabetes, with the figure anticipated to rise to 25% and 51% in 2030 and 2045, respectively[4].

Several diabetes prediction approaches have been proposed and published in recent years. Authors introduced ML-based method in which they used various dimensionality reduction and cross-validation techniques to apply Linear Discriminant Analysis[5], Quadratic Discriminant Analysis[6], Naive Bayes[7], Gaussian Process Classification[6], Support Vector Machine[8], AdaBoost[9], Logistic Regression[10], Decision Tree[11], and Random Forest. They also did a lot of testing on missing value filling and outlier rejection in order to increase the ML model's consistency.

The authors used three separate machine learning classifiers, such as Support Vector Machine, Naive Bayes and Decision Tree, to predict the probability of diabetes with the greatest accuracy. They got an accuracy of 0.819 and demonstrated that Naive Bayes is an excellent performing model[12].

Authors in [13] used four machine learning approaches to identify the probability of diabetes mellitus, namely, Logistic Regression, Naive Bayes and Decision Tree, with bagging and boosting strategies to improve robustness. The experimental results indicate that of all the algorithms used, the Random Forest algorithm produces the best results. Despite the fact that several frameworks have been released in recent years, there is still room for progress in diabetes prediction, precision and robustness.

A new procedure is being proposed for the prediction of diabetes using PIMA Indians Diabetes dataset in this paper. Outlier exclusion, missed value filling, function selection, and cross-validation are all part of the preprocessing process. In the missing location of an attribute, we choose the meanvalue instead of the median value since it favors the attribute's mean more centrally.In our planned pipeline, we used the ML classifiers Random Forest as well as XGBoost. We have applied the randomized search technique for selecting hyperparameters of XGBoost model and estimators for selecting the best creating a classifier in XGBoost.To choose the right parameters for our classifier, we used an estimator. In similar test conditions and dataset, extensive studies are carried out to improve diabetes prediction accuracy, researchers used various combinations of preprocessing and machine learning classifiers. The best ML model is then used as a starting point to construct a quantitative test for our proposed model for diabetes prediction precision. For improving diabetes prediction, we suggest a model based on a variety of ML models. To discover the best ensemble classifier, extensive testing on multiple types of ML models are conducted, which employs the best performing preprocessing from previous experiments.

2. METHODOLOGY

This segment reflects on the study's materials and processes.

(i) Dataset

The PIMA Indians Diabetes (PID) dataset was used to train and evaluate the models. It has data of Pima Indians community near Phoenix, Arizona, which has 768 female patients with diabetes [14]. Thereare500 non-diabetic patients (negative) and 268 diabetic patients (positive)and in this dataset, each with nine separate attributes. Table shows the attribute definitions as well as a short statistical overview.

Table 1: Overview of Diabetes *Patients*

S.No	Features	Mean ± std
1	Pregnant (F1)	3.84 ± 3.37
2	Glucose conc (F2)	120.89 ± 31.97
3	Blood Pressure (F3)	69.10 ± 19.36
4	Thickness (F4)	20.54 ± 15.95
5	Insulin (F5)	79.81 ± 115.24
6	BMI (F6)	31.99 ± 7.88
7	Pedigree (F7)	0.47 ± 0.33
8	Age (F8)	33.24 ± 11.76
9	Skin (F9)	0.81 ± 0.63

Pedigree Function (F7) is calculated as

$$F7= \frac{\Sigma_i k_i(88-ADM_i)+20}{\Sigma_j k_j(ALC_j-14)+50}$$

Where,i = developed diabetes

j = doesn't has diabetes

k = proportion of relatives who shares their gene

ADM_i = Relatives' ages at the time of the examination

ACC_j = Relatives' ages at the time of the non-diabetic examination

3. FRAMEWORK

Preprocessing of raw data is an important phase in o suggested pipeline in this literature since the consistency of data will motivate the classifiers to learn directly.

(i) Preprocessing

In our accepted system, the preprocessing step involves filling missed values (P), attribute function selection and standardisation (Q), which are defined as follows.

The attributes are processed to fill in empty or null values[15], which could cause classifiers to make incorrect predictions. Instead of dropping, the attribute mean values infers missing values in the accepted system, which can be formulated.

P(x) = mean(x), if x = missing

P(x) = x, else

Where, x = instance of feature vector

The technique of rescaling attributes to achieve a consistent normal distribution with zero mean and unit variance is known as standardisation or Z-score normalisation. The data distribution's skewness is also reduced by standardisation (Q).

$$Q(x) = \frac{x - \bar{x}}{\sigma}$$

Where, x = The feature vector's n-dimensional occurrences

$\bar{x}$ = mean of the attribute

σ = Standard deviation of the attribute

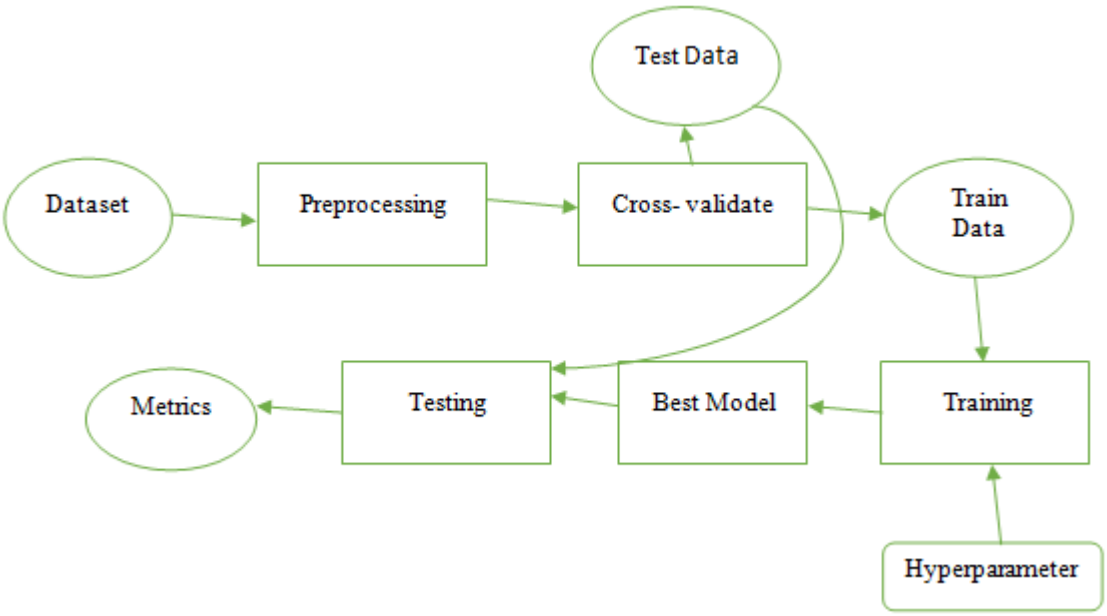

Figure 1: Flow Diagram of Prediction Model.

The classifiers' precision improves as the attribute's dimension is increased. When the attribute's value grows without the sample size growing, however, the classifiers' accuracy tends to deteriorate. The disadvantage of dimensionality is the term used in machine learning to describe such a situation. As a result of the dimensionality disadvantage, the function's space becomes increasingly sparse, forcing classifiers to be overfitted and lose generalising capabilities.

(ii) Cross Validation

The Cross-validation (CV) technique is perhaps the most popular common approaches for classifier error estimation and choosing a model among researchers[16]. This literature uses a pictorial representation of data slicing (5-fold cross-validation). The PID dataset is been divided into various number of folds for analysis. Grid search technique is used for the inner loop[17], the various number of folds are used for training and fine-tuning the hyperparameters. Test data and Optimum hyperparameters are utilized for validation of the model in the outer loop (K times). Due to the uneven negative and positive samples in PID dataset, split CV [30] was used to keep the same amount of samples per class as the initial percentage. An equation was used to calculate the final output metric.

$$R = \frac{1}{k} * \sum_{n=1}^{k} P_n \pm \sqrt{\frac{\sum_{n=1}^{k}(p - \bar{p})^2}{k-1}}$$

Where, R = Classifier's final performance metric

K = each fold's performance metrics.

(iii) ML model

The suggested model has been used to train and validate various machine learning models such as Decision Tree and XGBoost. In the inner circle, the hyperparameters that can tune.

4. METRICS OF EVALUATION

Models are created utilizing Python programming language and with KerasAPIs and various Python libraries. The tests were performed on a Windows 10 machine with the specified hardware setup: AMD Radeon (TM) R7 M360 GPU with 4GB GDDR3 900 MHz memory and Intel® CoreTM i5-6200 U CPU @ 2.30 GHz Dual Core processor with 8.0 GB (RAM).

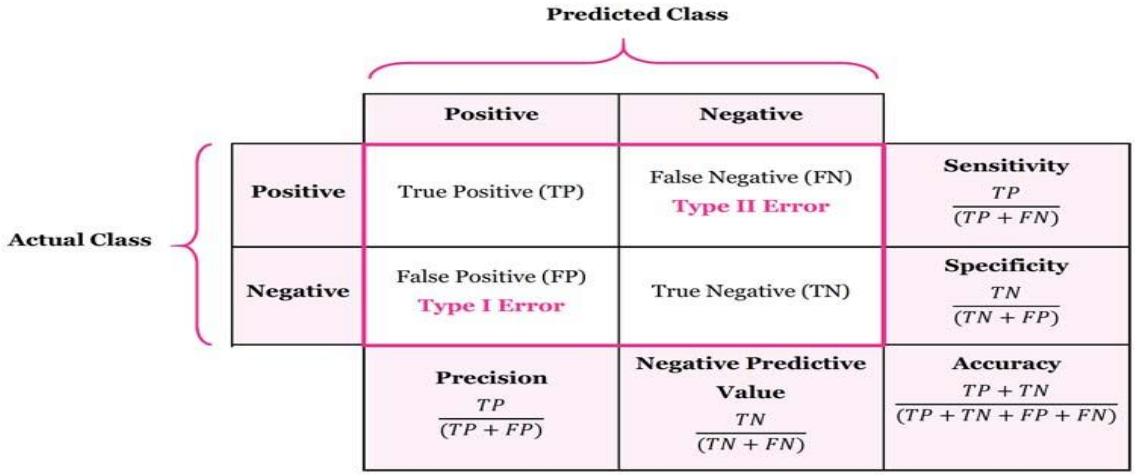

Figure 2: Confusion Matrix.

All of the detailed tests were analysed using a variety of criteria, each with its own sense of evaluation. There has been a study on the confusion matrix. The matrix of False Positive (FP), True Positive (TP), True Negative (TN), and False Negative (FN), as well as various parameters such as Precision (P), Specificity (Sp), False Omission Rate (FOR), Sensitivity (Sn), and Diagnostic Odds Ratio (DOR) [18]. The Sp and Sn are utilized to measure the type-II error (when a patient has positive symptoms and is incorrectly rejected), similarly type-I error (when a patient has negative symptoms and is incorrectly taken as positive).FOR, DOR and Pr has been utilized to assess the ratio of properly diagnosed diabetic patients with positive diagnoses, the proportion of people who have a negative test outcome but have a positive true diagnosis, and the diagnostic test's efficacy, respectively.

5. RESULT

This section divides the various detailed experiments into subsections, each with its own set of findings.

(i) Results for Preprocessing

The dataset's distribution based on class of attributes shows how difficult it is to differentiate between the different negative and positive diabetes. The inclusion of an outlier in the distribution of the feature adds kurtosis and skewness, with high kurtosis indicating outliers or heavy tails of PID dataset. Existence of skewness can cause the predicted values to be underestimated, while kutosis causes the predicted value to be overestimated. The outlier exclusion outcome suggests that the distribution's skewness changes to zero mean, showing that the attribute's median and mean have nearly interacted.

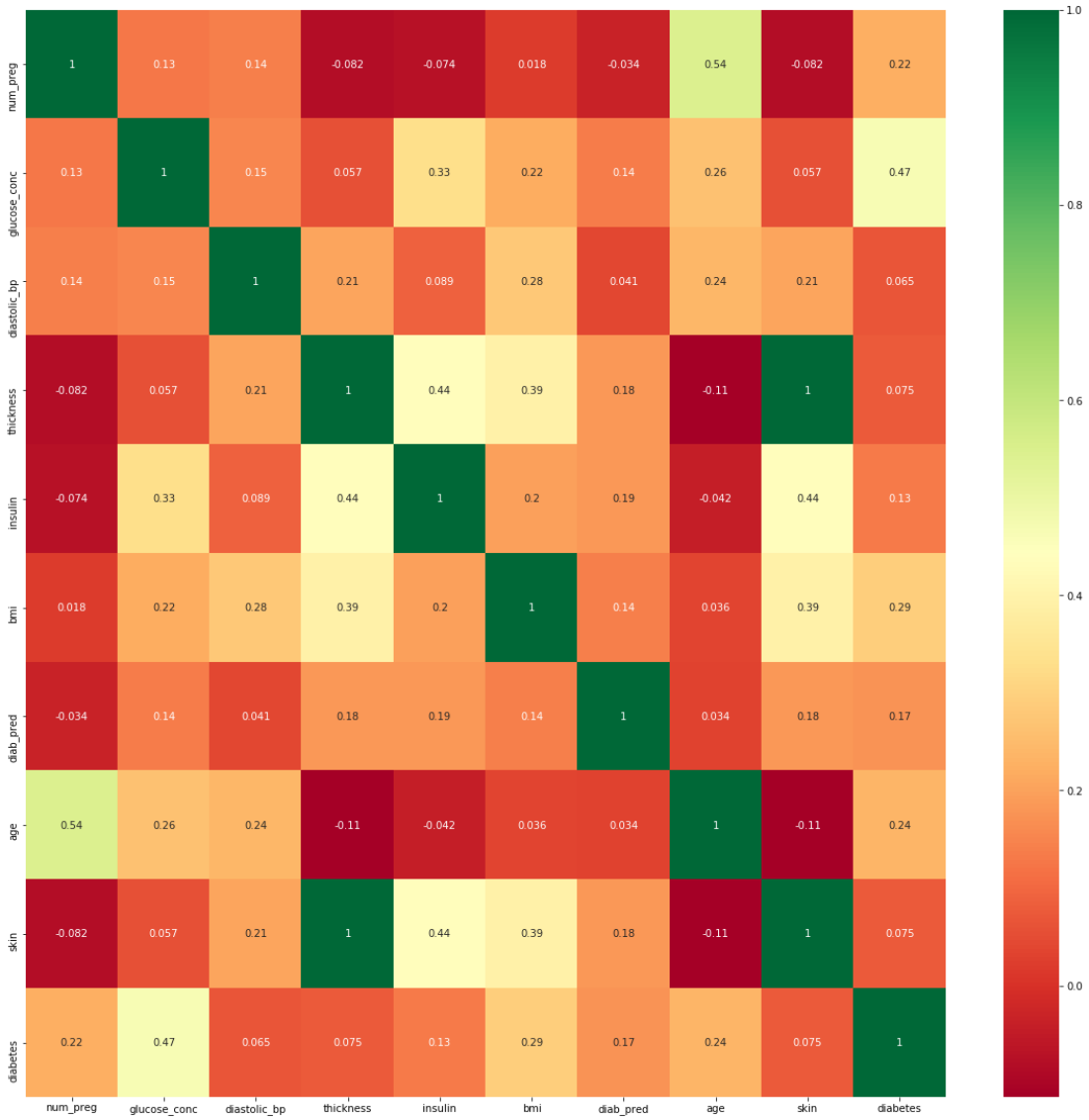

Figure 1: The Attribute's Connection with The Outcome Is Represented By A Confusion Matrix.

(ii)Results for ML Model

Research results of selecting the best performing preprocessing and machine learning model, with accuracy and standard deviation recorded for comparison in table. Table shows an overview of each model's ability to achieve the best accuracy from the accepted pipeline, along with the most effective preprocessing with attribute collection algorithms and the number of chosen attributes. Table 3 also includes the most effectively optimized hyperparameter obtained by grid search. Table 3 shows that when we use enough preprocessing, we can get better outcomes from various models.

Table 2: Summary of all Tests for Determining the Highest Performing Preprocessing, Feature Selection Processes, and Classifier Using Chosen Attribute Numbers

Preprocessing	Algorithm	Random Forest	XGBoost	Best
Raw Data	N/A	0.829 ± 0.035	0.832 ± 0.034	XGBoost
Outlier Rejection	N/A	0.831 ± 0.039	0.836 ± 0.039	XGBoost
Outlier Rejection	Correlation based feature selection	0.881 ± 0.035	0.885 ± 0.049	XGBoost
Outlier Rejection + Filling missing value with mean	N/A	0.937 ± 0.015	0.941 ± 0.024	XGBoost
Outlier Rejection + Filling missing value with mean	Correlation based feature selection	0.944 ± 0.027	0.951 ± 0.028	XGBoost

The correlation between the feature and the desired result is used to pick the features in correlation-based feature selection. When the correlation-based feature selection is used, both of the classifiers provide their best performance for filling missed values. As seen in Table 3, boosting classifiers (XGBoost) outperform Random Forest classifiers in terms of accuracy. In the PID dataset, the XB performs better for both raw as well as for preprocessed data.DespiteXGBoost's extreme gradient boosting capabilities, these experiments reveal that XGBoost is more influenced based on the PID dataset's outlier than Random Forest. There is a risk of overfitting. For the preprocessing of P, the XGBoost outperforms the Random Forest classifier for correlation-based feature collection. Since the outcome is associated with characteristics from the correlation based selection.

Classification accuracy has improved greatly when instead of rejection, missing values are replaced with the mean. When both the P and Q are used, the XGBoost has prevailed in any case of function selection. Both the classifiers perform admirably in the pre-process (P + Q) and correlation-based function collection, with no missing values or outliers. When comparing both the ML models in Tables 3 and 4, the XGBoost has the highest accuracy of 0.951 ± 0.028, due to its extreme gradient boosting is a feature that can help to avoid failure while adding new models.

Table 3: Most Effective ML Model With Preprocessing and Optimized Hyper Parameters

Ml Model	Preprocessing	Hyperparameters	Best Performance
Random Forest	Correlation based feature selection + Filling missing value with mean	criterion = gini random_state = 10 n_estimator =100	0.944 ± 0.027
XGBoost	Correlation based feature selection + Filling missing value with mean	gamma = 0.5 min_child_weight = 1 max_depth = 4 subsample= 1.0 Learning_rate = 0.2 colssample_bytree =0.7	0.951 ± 0.028

REFERENCES

1. *Centers for Disease Control and Prevention website," 18 July 2017. [Online]. Available: www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf. [Accessed 1 August 2017].*

2. *T. E. R. F. Collaboration, "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies," The Lancet, vol. 375, no. 9733, pp. 2215-2222, 26 June 2010.*
3. *J. E. S. S. K. Y. H. J. D. R. F. A. W. O. a. B. M. N. H. Choac, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," Diabetes Research and Clinical Practice, vol. 138, pp. 271-281, April 2018.*
4. *I. P. P. S. B. M. S. K. N. U. S. C. L. G. A. A. M. K. O. J. E. S. D. B. R. W. a. I. D. A. C. P. Saeedi, "Goba and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045," Diabetes Research and Clinical Practice, vol. 157, pp. 107-843, November 2019.*
5. *R. K. Gaurav Tripathi, "Early Prediction of Diabetes Mellitus Using Machine Learning," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), pp. 1009-1014, June 2020.*
6. *K. M. K. M. T. T. Z. M. Emon MU, "Primary Stage of Diabetes Prediction using Machine Learning Approaches," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 364-367, March 2021.*
7. *K. L. K. M. S. C. R. R. M. M. S. &. R. G. R. M. Priya, "A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 603-607, June 2020.*
8. *N. a. V. J. Mohan, "Performance Analysis of Support Vector Machine in Diabetes Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, pp. 1-3, November 2020.*
9. *N. H. A. I. a. I. M. Taz, "A Comparative Analysis of Ensemble Based Machine Learning Techniques for Diabetes Identification," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 1-6, Jan 2021.*
10. *H. H. B. T. a. A. A. S. Alshamlan, "A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression," 2020 11th International Conference on Information and Communication Systems (ICICS), pp. 1-4, 7 April 2020.*
11. *A. M. S. V. a. D. J. R. Posonia, "Machine Learning based Diabetes Prediction using Decision Tree J48," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 498-502, December 2020.*
12. *D. S. a. D. S. Sisodia, "Prediction of diabetes using classification," Procedia Computer Science, vol. 132, pp. 1578-1585, January 2018.*
13. *N. N.-a. a. R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," Procedia Computer Science, vol. 69, pp. 132-142, December 2015.*
14. *J. E. E. W. C. D. W. C. K. a. R. S. J. J. W. Smith, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in Proc. Annual Symposium on Computer Application in Medical Care, pp. 261-265, November 1988.*
15. *D. C. a. S. Chartier, "Outliers detection and treatment: A review," International Journal of Psychological Research, vol. 3, no. 1, pp. 58-67, March 2010.*
16. *S. A. a. A. Celisse, "A survey of cross-validation procedures for model selection," Statistics surveys, vol. 4, pp. 40-79, July 2010.*
17. *L. B. D. E. L. a. S. T. D. Krstajic, "Cross-validation pitfalls when selecting and assessing regression and classification models," Journal of Cheminformatics, vol. 6, no. 1, p. 10, March 2014.*
18. *J. G. L. M. H. P. G. J. B. a. P. M. M. B. A. S. Glas, "The Diagnostic Odds Ratio: A single indicator of test performance," Journal of Clinical Epidemiology, vol. 56, no. 11, pp. 1129-1135, November 2003.*

19. *S. A. a. A. Celisse, "A survey of cross-validation procedures for model selection," " Statistics surveys, vol. 4, pp. 40-79, July 2010.*
20. *Abbas, Hamed Jadooa, and Jawad Mohammad Ismail. "Association of Adiponectin Levels in Iraqi Patients With Type 2 Diabetic and Prediction of Cardiovascular Complications." Tjprc: International Journal of Diabetes & Research (Tjprc: Ijdr) 2.1 (2016) 1 8 (2016).*
21. *Gunjigavi, Sanjeev Kumar S., T. Anil Kumar, and Ashwin Kulkarni. "Study Of Ischemic Heart Disease Among Patients With Asymptomatic Type-2 Diabetes Mellitus In A Tertiary Hospital In South India Using Computed Tomographic Angiography." International Journal of Medicine and Pharmaceutical Sciences (IJMPS) 10 (2020): 9-18.*
22. *Al-Naama, Lamia Mustafa, And Jawad Kadham Mahdi. "The Role Of Islet Cell Autoantibodies, Islet Cell Antigen-2 Antibody And Antioxidant Enzymes In Diabetic Patients Type 2. International Journal Of Medicine And Pharmaceutical Science (Ijmps) 6.1, Feb 2016, 37-46.*
23. *Thi-Qar, I. I. N. "Investigate The Relation Between Ctla4 Gene Polymorphisms And Insulin Dependent Diabetes Mellitus (Iddm) Type I In Thi-Qar Population." International Journal Of Medicine And Pharmaceutical Sciences (Ijmps) 4.6, Dec 2014, 45-54*